A Safety Detection Model for Substation Operations with Fused Contextual Information

Chen Bo[1], Zhanghong Yu[2], Yangrun Xi[3], Zhao Lei[4], Ding Yi[5*]
State Grid Beijing Electric Power Company, Beijing 100031, China[1, 2, 3, 4]
Beijing Electric Power Economic Research Institute Co., Ltd., Beijing 100055, China[1, 2, 3, 4]
Nanjing Artificial Intelligence Research of IA, Nanjing 211100, China[5]

Abstract—**Detecting and regulating compliance at substation construction sites is critical to ensure the safety of workers. The complex backgrounds and diverse scenes of construction sites, as well as the variations in camera angles and distances, make the object detection models face low accuracy and missed detection problems. In addition, the high complexity of existing models creates an urgent need for effective parameter compression techniques to facilitate deployment at the edge server. To cope with these challenges, this study proposes a safety protection detection algorithm that fuses contextual information for substation operation sites, which enhances multi-scale feature learning through a two-path downsampling (TPD) module to effectively cope with changes in target scales. Meanwhile, the Global and Local Context Information extraction (GLCI) module is utilized to optimize the key information learning and reduce the background interference. Furthermore, the C3GhostNetV2 unit is utilized in discerning the interconnections of far-off spatial pixels, while enhancing the network's expressive power and reducing the number of parameters and computational costs. The outcomes of the experiments indicate that the present model improves upon the mAP50 metric by 4.5% compared to the baseline model, and the accuracy of the checks and the recall have seen respective increases of 4.8% and 10.1%, while there has been a reduction in both the count of parameters and the floating-point calculations by 16.5% and 12.6% respectively, which proves the validity and practicability of the method.**

Keywords—Object detection; context information; electricity construction operation; model complexity; lightweight

I. INTRODUCTION

As societal demand for electrical energy continues to rise, there is an increasingly urgent need for power production in China. However, due to a combination of various factors, the frequency of power production accidents in the country remains relatively high, posing a serious threat to urban safety. The power construction process involves a wide range of complex scenarios, including tower assembly, hoisting, excavation work, hot work, edge work, and high-altitude operations. These tasks encompass numerous safety challenges, requiring workers to maintain a high level of vigilance and adhere to standardized procedures to prevent serious accidents [1-2].To ensure that power workers enhance their safety awareness, adopt compliant protective measures and procedures, and ultimately improve on-site safety levels to ensure the normal operation of the power system, there is an urgent need for worksite compliance monitoring and supervision. Traditional manual management methods are costly and inefficient, making it difficult to meet the needs of effective supervision in multi-scenario, around-the-clock power grid operations. Therefore, the application of deep learning algorithms for compliance monitoring at power grid work sites holds significant research value. This research introduces advanced visual technologies to the field of power production and provides substantial support for improving on-site safety levels.

With the tremendous achievements of deep learning algorithms in image recognition and detection [3-5], these algorithms have made significant progress in various applications. Thanks to their high detection accuracy and strong robustness [6-7], deep learning-based object detection algorithms have been widely applied in detecting standardized work clothing. Ren et al. [8] discussed the concept of deep learning-based intelligent substation system monitoring and analyzed the advantages and disadvantages of using traditional methods and deep learning for monitoring. Liu et al. [9] employed the Faster R-CNN algorithm for detecting whether standard work clothing is worn and introduced an L2 regularization term into the loss function to improve the convergence speed of the model during training. The model demonstrated good generalization ability and robustness, with significant improvements in both accuracy and real-time performance compared to the baseline model. Reference [10] employed the lightweight network from MobileNet in the YOLOv2 structure to achieve a certain degree of network compression, reducing the computational complexity of the model and improving its convergence performance, but with lower accuracy in object detection. Xu et al. [11] proposed an improved YOLOv3 algorithm for safety helmet recognition, which enhanced the precision of safety helmet detection, but the detection speed was relatively slow. The study in [12] implemented a real-time video analysis algorithm based on YOLOv4 for monitoring whether workers in industrial facilities wear helmets, safety vests, and safety belts, but it ignored the influence of complex backgrounds and environmental factors on algorithm performance and did not analyze the algorithm's complexity. Du et al. [13] selected the Swin Transformer as the backbone network based on YOLOv5 to extract deeper semantic information and capture more detailed features of safety helmets, but it had false detection issues when the colors were the same. In addition to YOLO, Long et al. [14], Wu et al. [15], and Li et al. [16] proposed safety helmet detection methods based on SSD, which achieved good detection results. Although single-stage object detection algorithms have performed well in terms of real-time performance and efficiency, they still face some challenges, such as the use of dense grids or anchor boxes to generate candidate regions, which can lead to overlapping or missed detections and relatively poor performance in detecting small objects [17].

Existing detection methods exhibit certain limitations in practical application scenarios. They often focus on specific target detection while neglecting the diverse task requirements in power construction sites. For instance, Shen et al. [18] utilized convolutional neural networks to detect facial features and helmet usage on construction sites. However, this study did not sufficiently address the impact of environmental variations on detection performance. Additionally, the lightweight improvement of YOLOv5 proposed in [19] targets helmet detection only and still has room for improvement in adaptability and scalability in real-world scenarios. Furthermore, the methods in [20, 21] are only effective in environments with simple backgrounds. These approaches, which focus on single categories or specific scenarios, struggle to handle the complexity and variability of power construction environments, limiting their broader applicability in practical settings.

In recent years, research on compliance detection for multiple categories of personal protective equipment (PPE) has increased. For instance, Zhang et al. [22] enhanced YOLOv4's feature extraction network and combined it with PANet to achieve effective detection of helmets and masks. Similarly, [23] utilized synthetic datasets to train an improved YOLOv5 model, successfully applying it to real-time PPE detection in industrial environments. In study [24], an intelligent detection system was designed to notify supervisors when workers failed to wear helmets or vests, improving the efficiency of on-site safety management. Additionally, Gong et al. [25] adopted a key region localization method to optimize PPE detection performance further. While these methods have achieved notable advances in accuracy and detection capability, they still face certain limitations. The complex model structures demand significant hardware resources and result in low detection efficiency, making them challenging to apply in industrial scenarios requiring real-time performance and low power consumption. Therefore, reducing model complexity while maintaining detection accuracy remains a critical research direction in this field.

The aforementioned methods have achieved detection and monitoring of power grid operation scenes with humans as the main objects to some extent, but the detection scenarios and categories are relatively limited, usually focusing on specific categories such as safety helmets or work clothing, with less attention to other scenes. This limits the comprehensiveness and applicability of the algorithms because power grid operation scenes involve various scenarios such as lifting operations, excavation work, hot work, work near edges, and work at heights, requiring more comprehensive detection capabilities. Additionally, the collected images from the power grid industry exhibit characteristics of diverse target types, inconsistent sizes, and complex and variable background environments. In complex backgrounds, the previous algorithms perform poorly in effectively extracting features from targets of different scales, resulting in weak adaptability to environmental changes. This can lead to issues of missed detections and low accuracy in practical applications, reducing the reliability of the algorithms. Furthermore, existing models have high complexity, requiring a reasonable and effective parameter simplification technique as a basis to address the challenges of deployment on remote server devices. Enhancing object detection performance in complex scenarios, improving the detection accuracy of multi-scale objects, and reducing model complexity have become key research focuses in substation safety monitoring. Therefore, this study presents an improved YOLOv7 network. A down-sampling module was designed to enable the model to better learn multi-scale features. Additionally, a Global and Local Context Information extraction (GLCI) module was introduced, allowing the model to effectively capture critical information within complex backgrounds and reduce missed and false detections caused by background interference. Furthermore, structural optimization was implemented to reduce the model’s parameter count, enhancing the real-time performance and adaptability of the improved YOLOv7 model while maintaining high detection accuracy. These improvements enable efficient and compliant detection in power production scenarios. Our contributions are summarized as follows:

- We design a global and local context information extraction (GLCI) module, enabling the network to capture both global contextual and local spatial information, effectively addressing the challenge of complex backgrounds in power construction site environments.
- We propose a two path downsampling (TPD) module, which enhances the network’s ability to learn features across multiple scales, improving performance on multi-scale target detection tasks.
- We develop a novel C3GhostNetV2 unit, replacing all ELAN-H modules in the neck network and the ELAN modules at the backbone’s end. This design expands the receptive field, strengthens model representation, and significantly reduces model complexity, parameter count, and computational cost.

II. Proposed Method

This method consists of three modules: Global and Local Context Information extraction (GLCI) module Two-Path Downsample (TPD) module, and C3GhostNetV2 module. The TPD module enables the network to effectively capture and utilize spatial information from feature maps of different scales. The GLCI module helps the network learn key information more efficiently and reduces background interference. The C3GhostNetV2 unit not only expands the perception range to ensure the model's expressive effectiveness but also reduces the demand for floating-point calculations.

The YOLO series algorithms have high similarity in terms of network structure and module composition. Taking YOLOv7 as an example, Fig. 1 shows the network structure diagram of YOLOv7 with GLCI and TPD inserted and lightweight improvements applied. It mainly includes four components: Input, Backbone, Neck, and Head. The Backbone consists of CBS, C3, and SPPF, which are responsible for extracting information features from input images. After the Backbone, feature maps of three different sizes can be obtained. The Neck module combines the Feature Pyramid Network (FPN) and Pyramid Attention Network (PAN) to fuse the features extracted by the Backbone. The Prediction Head consists of three prediction layers of different scales, which are responsible for outputting the network's predictions.

In YOLO series networks, the backbone is primarily responsible for feature extraction, transforming input images into high-level feature representations [26], which serve as the basis for subsequent detection. The insertion of the GLCI module enhances the network's learning ability for multi-scale information and its perception of global and local context, addressing the challenges posed by significant scale variations and complex backgrounds. The backbone is where the network learns and perceives information, making it a suitable location for the insertion of the GLCI module. Additionally, placing the GLCI plugin at the end of the backbone allows for effective utilization of high-level information in the feature maps, and the lower resolution of the end feature maps can alleviate the potential computational and memory costs introduced by the plugin. Furthermore, since the proposed GLCI plugin does not change the size of the input and output feature maps, it is not affected by the number of network layers and is applicable to all YOLO series algorithms. Taking YOLOv7 algorithm as an example, as shown in Fig. 1, the feature maps output by the last CBS module in the backbone have a size of 20×20×1024. After being processed by the GLCI plugin, the size and number of channels of the feature maps remain unchanged, meeting the data format requirements of the subsequent network structure. Therefore, it is reasonable to insert the GLCI module at the end of the backbone. Considering that the neck network contains rich information and serves as the direct input to the head network, the TPD module is inserted into the neck network. Finally, the C3GhostNetV2 unit is constructed in the neck network to expand the perception field and maintain the model's expressive power while significantly reducing the number of parameters and computational costs.

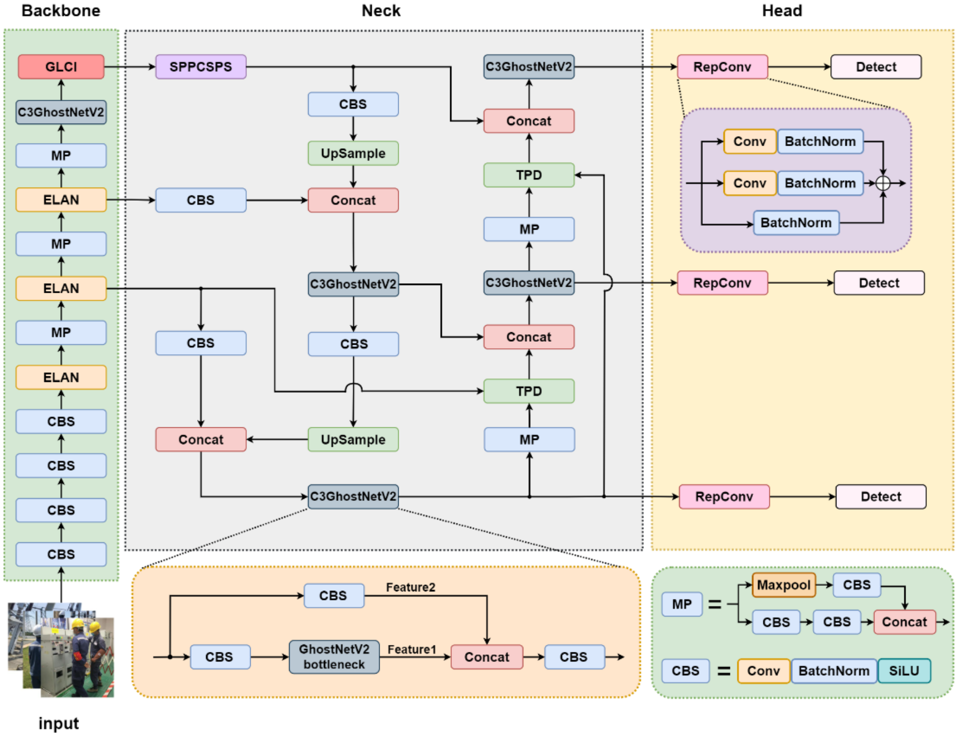

Fig. 1. Schematic diagram of lightweighting methods structure.

A. *Global and Local Context Information Extraction Module*

To improve network adaptability in complex image backgrounds in real-world power construction site detection tasks, we propose the GLCI module (see Fig. 2). This module consists of two branches: Global and Local Context Extraction. It helps the object detection network learn more crucial information and attenuate background interference. The primary objects to be detected in input images—power construction site workers—are closely tied to their surrounding environments. Incorporating contextual information into the model enhances its understanding of the relationships between detected objects and their scenes, thereby improving detection performance. The global branch of the GLCI module, built on the traditional self-attention mechanism, leverages not only the relationships between keys and queries but also emphasizes the contextual information among input keys. This approach enhances the network's capacity to extract contextual information and improves its ability to learn critical features through the guidance of the learned dynamic attention matrix.

For a feature map X of size $c \times h \times w$, linear processing is applied to yield K=X, Q=X,, and V=XWV, where K denotes the key, Q the query, and V the value, with WV representing a 1×1 embedding convolution. In the spatial domain, group convolution is performed on adjacent keys within a 3×3 grid of ***K***, with the number of groups set to 4. This is followed by batch normalization (BN) and ReLU activation, resulting in a feature map $\boldsymbol{K_1}$ of size $c \times h \times w$. Through these operations, encoding is applied to the adjacent keys in the spatial domain, producing K1, which captures the static contextual information between neighboring keys and is referred to as static contextual keys $\boldsymbol{K}_1$. Subsequently, $\boldsymbol{K}_1$ and $\boldsymbol{Q}$ are concatenated, and two successive 1×1 convolutions generate an attention matrix $\boldsymbol{A}$. This attention matrix $\boldsymbol{A}$ differs from the traditional self-attention mechanism, as it is derived from query features and static contextual features of $\boldsymbol{k}_1$, rather than key/query pairs. Thus, $\boldsymbol{A}$ effectively aggregates contextual information.

$$A = [K^1, Q]W_\theta W_\delta \tag{1}$$

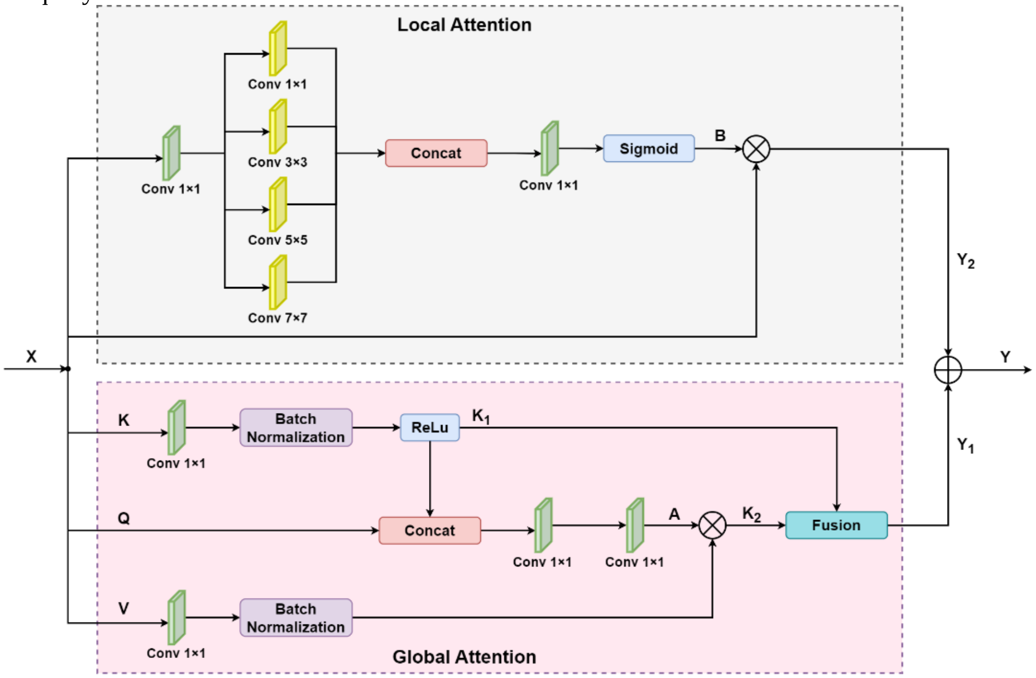

Fig. 2. Structure diagram of GLCI module.

In Eq. (1), W_θ represents a convolution with a ReLU activation function, and W_δ represents a convolution without an activation function. Next, the attention matrix $\boldsymbol{A}$, which aggregates contextual information, is element-wise multiplied with $\boldsymbol{V}$ resulting in a feature map $\boldsymbol{K}_2$ weighted by $\boldsymbol{V}$. Since $\boldsymbol{K}_2$ is obtained through self-attention computation on the input keys and values, it captures the dynamic feature interactions among the inputs and is named the dynamic contextual keys. The fusion of the static contextual keys $\boldsymbol{K}_1$ and the dynamic contextual keys $\boldsymbol{K}_2$ yields the output $\boldsymbol{Y}_1$ of the global attention network.

In the Local Contextual Information (LCI) network, multiple convolution sizes capture spatial information at various scales, enhancing learning and addressing noisy, complex backgrounds. For an input feature map $\boldsymbol{X}$ of size $c \times h \times w$, a 1×1 convolution reduces the channels to 1/16, minimizing parameters. The reduced map is processed by 1×1, 3×3, 5×5, and 7×7 convolutions, concatenated into a tensor of size $1/4\,c \times h \times w$, then passed through a 1×1 convolution to reduce the channels to 1. Applying the sigmoid function produces a local spatial attention map $\boldsymbol{B}$ of size 1×H×W, which is element-wise multiplied with $\boldsymbol{X}$ to generate $\boldsymbol{Y}_2$. The outputs of the Global and Local Contextual Information networks, $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$, are summed to produce $\boldsymbol{Y}$, enabling the GLCI module to capture both local and global features and address challenges from complex backgrounds.

B. Two Path Downsampling (TPD) Module

The challenge of handling multiscale data in deep learning lies in efficiently extracting and learning features from data of varying scales, such as different sizes and resolutions. To address the impact of multiscale issues on detection accuracy in images collected from power construction sites, this paper proposes the Two Path Downsampling (TPD) module. This module facilitates information exchange between different feature layers, enabling spatial channel dependencies between different scales and enhancing the performance of feature extraction across different scales. The TPD module's structure is illustrated in Fig. 3.

The TPD module takes two input feature maps: the local feature map F_c with dimensions $c \times h \times w$ and the higher-level feature map F_u with dimensions $1/2c \times 2h \times 2w$. Unlike traditional stride-based convolutional downsampling, this module introduces a lossless downsampling process for the local feature map, preserving its fine-grained details. Additionally, the downsampling process is improved by preserving spatial information before and after downsampling, reducing the loss of detail. Furthermore, the module captures details from the higher-level feature map and incorporates semantic information from the local small-scale feature map, enhancing interdependencies between feature maps at different levels.

The downsampling process consists of two distinct paths. Path 1 involves lossless downsampling of the local feature and an enhanced convolutional downsampling process. The local feature F_c first undergoes feature extraction through average pooling and convolution with stride 1, resulting in the feature map F_{cl} of size $c \times h \times w$. Then, the feature map F_{cl} is reshaped to generate the pseudo lower-level feature F_{nl} of size $c\times 1/2h\times 1/2w\times 4$, which preserves the local feature without information loss. Subsequently, softmax normalization is applied to obtain the lossless downsampled result F_{snl} of size $c\times 1/2h\times 1/2w\times 4$. Meanwhile, the spatial information of the local feature F_c is preserved by applying a convolution operation with stride 1, resulting in the spatial information feature F_s of size $4c\times h\times w$. Then, a downsampling operation is performed using a convolutional operation with stride 2 to generate the lower-level feature F_{sl} of size $4c\times 1/2h\times 1/2w$,which includes the preserved spatial information. Next, the feature F_{sl} undergoes feature extraction using the BAM [27] attention mechanism, yielding the feature representation F_{bsl} . Finally, the previously preserved spatial information is restored through a reshaping operation, generating the lower-level feature representation F_{el} of size $c\times 1/2h\times 1/2w\times 4$, which includes enhanced spatial information.

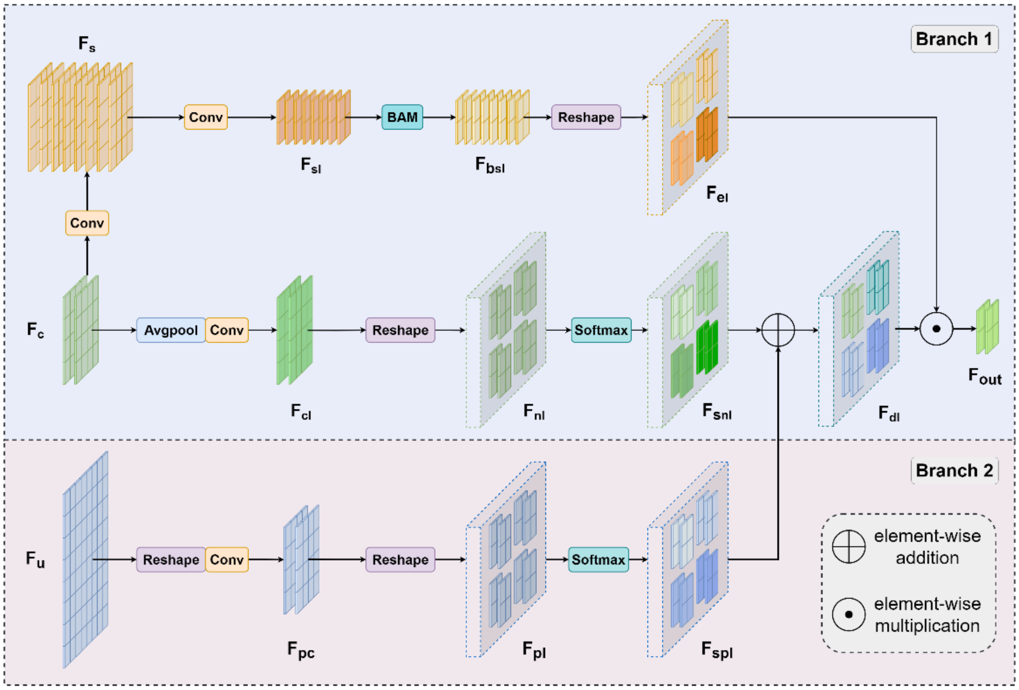

Fig. 3. Structure diagram of TPD module.

Path 2 involves a lossless downsampling process for the higher-level feature. The higher-level feature undergoes downsampling for the first time through reshaping, resulting in a size of $2c\times h\times w$. Then, a convolution operation is applied to halve the number of channels, generating a pseudo local-level feature F_{pc} based on the reconstructed details from the higher-level feature. Subsequently, a second downsampling is performed, yielding a pseudo lower-level feature F_{pl} based on the reconstructed details from the higher-level feature, with a size of $c\times 1/2h\times 1/2w\times 4$. Finally, softmax normalization is applied across the last dimension to obtain the output F_{spl} of Path 2.

The output F_{spl} of Path 2 is element-wise added with the downsampling result F_{snl} of the local feature, resulting in a lower-level feature map of size $c\times 1/2h\times 1/2w\times 4$ that combines the detailed information from both the higher-level and local features. This lower-level feature, denoted as F_{dl} , contains the detailed information from both the higher-level and local feature maps. The equation is as follows:

$$F_{dl} = Softmax(R(Conv(R(F_u)))) + \\ Softmax(R(Conv(Avgpool(F_c)))) \tag{2}$$

Here, $R(\square)$ represents the reshape operation. The lower-level feature map F_{dl} obtained from the fusion in (2) is multiplied and fused with the enhanced lower-level feature F_{el} from Path 1, followed by summation across the last dimension. This yields the output feature map F_{out} of the TPD module, with a size of $c\times 1/2h\times 1/2w$. The equation is as follows:

$$F_{el} = R(BAM(Conv(Conv(F_c)))) \tag{3}$$

$$F_{out} = Sum(F_{dl} \square F_{el}) \tag{4}$$

C. C3GhostNetV2 Module

The original YOLOv7 model has relatively high complexity due to deep layers and multiple convolution operations, leading to many parameters and computational redundancy [28]. To reduce floating-point operations and parameters, this paper introduces the C3GhostNetV2 module. As shown in Fig. 1, the input feature map is split into two branches: one passes through CBS and the GhostNetV2 bottleneck[29], generating Feature 1, while the other passes through CBS to produce Feature 2. Feature 1 and Feature 2 are then concatenated and processed by CBS to produce the final output.

Fig. 4 illustrates the structure of the GhostNetV2 bottleneck, comprising three steps: First, the input feature map is transformed into a compressed low-dimensional vector through downsampling and convolution. Next, this vector is processed by fully connected layers in vertical and horizontal directions, expanding its receptive field across multiple dimensions. Finally, the attention weights are normalized using the Sigmoid activation function to enhance the network's utility and stability. As a result, the network is able to perceive long-range dependencies between spatial pixels, enhancing the expressive power of the model. The DFC attention output [30] is combined with the first Ghost module's output. Depth-wise separable convolution is used to further reduce computational and memory overhead, improving inference speed. After generating features from the second Ghost unit, a skip connection merges the initial input with the new features, producing the final output. This design captures long-range spatial dependencies while significantly reducing computational and parameter costs.

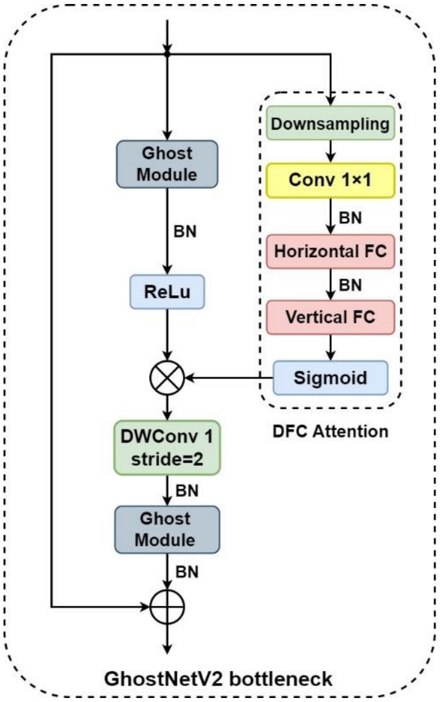

Fig. 4. GhostNetV2 bottleneck structure.

III. Experimental Preparations

A. Experimental Setup and Dataset

The software environment for this experiment includes Ubuntu 16.04, PyTorch 1.11, and CUDA 11.3. The hardware setup is described in detail in Table I. The ablation experiments were conducted using the stochastic gradient descent (SGD) strategy with 100 epochs of training. The initial learning rate was set to 0.01 and decayed with a minimum value of 0.0001. The batch size was set to 8. The momentum parameter was set to 0.9, and the weight decay was set to 0.0005.

TABLE I. Experimental Hardware Setup

Hardware Name	Model	Quantity
CPU	Intel Core i7-10700 CPU	1
Memory	Kingston 16G DDR4	2
Graphics Card	NVIDIA RTX-3090	1
Hard Drive	Western Digital 10TB	1

The experimental dataset used in this study consists of 4,030 images collected by a power company. The dataset includes four different work scenarios: lifting operations (1,104 images), hot work operations (1,028 images), edge operations (973 images), and high-altitude operations (925 images). These scenarios cover samples with different target sizes and brightness levels (see Fig. 5).

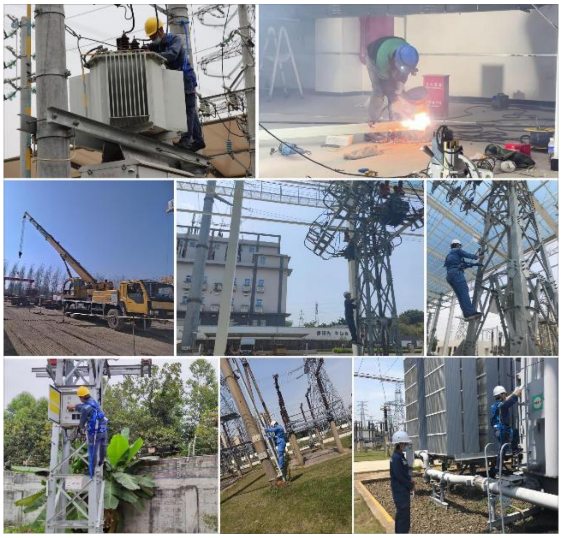

Fig. 5. Examples of 18 types of detection target datasets.

These work scenario data consist of eighteen categories: (1) crane, (2) all personnel on site, (3) wearing safety helmets, (4) work barriers, (5) control room, (6) hooks, (7) wearing safety harnesses, (8) not wearing safety harnesses, (9) safety harnesses properly suspended, (10) safety harnesses improperly suspended, (11) personnel performing hot work, (12) supervisors overseeing hot work, (13) protective face shields, (14) ignition source, (15) fire extinguisher, (16) improper wearing of safety gloves, (17) mobile phones. The dataset contains a total of 96,419 instances, which reflects the complexity and diversity of the work scenarios. The training set and validation set are split in an 8:2 ratio. Evaluation metrics

For the evaluation of the performance of the object detection model, a specific evaluation system was adopted. In this paper, the model's floating-point operations (GFLOPs) and the total number of parameters were calculated to assess its runtime and memory requirements. Meanwhile, the average precision (AP) and mean average precision (mAP) were used as standards to measure the accuracy of the model. The detection efficiency of the model on different categories was determined by the average precision rate, which is comprehensively determined by the recall and precision.

The recall is calculated using the following formula:

$$R_{ec} = \frac{T_p}{T_p + F_N} \times 100\% \quad (5)$$

The precision is calculated using the following formula:

$$P_{re} = \frac{T_p}{T_p + F_p} \times 100\% \quad (6)$$

Where T_p represents the true positive, F_N represents the false negative, and F_p represents the false positive. Precision-recall (PR) curve plots recall on the x-axis and maximum precision on the y-axis. The area under the PR curve is calculated by integrating over the curve, resulting in the value of AP (average precision). The mean average precision (mAP) is obtained by calculating the average of the AP values for all individual classes. The calculation formula is as follows:

$$AP = \int_0^1 P(r)dr \quad (7)$$

$$mAP = \frac{\sum_{k=0}^{c} AP_k}{C} \quad (8)$$

In the formula, $P(r)$ represents the PR curve, $\sum_{k=0}^{c} AP_k$ represents the average precision for each class, and C represents the total number of classes.

IV. Experimental Results And Analysis

A. Ablation Study

To evaluate the performance of the proposed approach in object detection tasks and study the effectiveness of various optimization strategies, we conducted ablation experiments on the YOLOv7 model as a baseline. In the table, the symbol " √ " indicates that the corresponding optimization unit has been applied. All experimental groups were conducted with the same hyperparameters and training strategies to analyze the impact of optimization strategies on the network more clearly. The experimental results are shown in Table II.

Experiment 1 was conducted on the original YOLOv7 without any improvements. In Experiment 2, the GLCI module was added. In this case, the precision did not change much, but the recall increased by 7.2%, and the mean average precision (mAP50) improved by 2.1%. This indicates that the GLCI module has a significant advantage in handling object detection tasks with complex backgrounds. By applying global and local attention mechanisms, the network learns key information more efficiently and reduces background interference. In Experiment 3, the TPD module was added. In this case, the precision increased by 2.5%, the recall increased by 8.7%, and mAP50 improved by 2.3%. This suggests that the introduction of the TPD module effectively enhances the network's learning ability for feature maps of different scales. By extracting and fusing multi-scale spatial and channel information, it effectively solves the problem of detecting objects with multi-scale variations and improves detection performance. From the results of Experiments 2 and 3, it can be seen that after integrating the TPD and GLCI modules, there is no significant difference in the number of parameters and the scale of floating-point operations. The introduction of the C3GhostNetV2 module in Experiment 4 resulted in a significant reduction of 87.2% in the scale of floating-point operations and a decrease in the parameter scale to 83.0%, effectively reducing the complexity of the model. The mAP50 metric of the model also increased by 0.4%, with corresponding improvements of 1.7% in precision and 9.6% in recall. This indicates that even with a reduction in parameters and computational complexity, the detection accuracy of the model was not compromised. This confirms that the C3GhostNetV2 unit not only reduces the complexity of the model but also enhances its performance. When all three modules (GLCI, TPD, and C3GhostNetV2) were simultaneously inserted into YOLOv7, the network showed the best improvement. The average precision (mAP50) increased by 4.5% compared to the original model, and precision and recall improved by 4.8% and 10.1% respectively. Moreover, compared to the baseline model, the complexity of the model was also reduced to a certain extent. This demonstrates that the simultaneous use of the three proposed improvement methods can yield better results.

TABLE II. Improved YOLOv7 Algorithm Ablation Study Results

Experiment	GLCI	TPD	C3GhostNetV2	mAP50	Precision	Recall	Parameters	GFLOPs
				/%	/%	/%	/M	/G
1				86.3	85.3	78.5	37.6	107.3
2	√			88.4	85.2	85.7	37.6	107.6
3		√		88.6	87.8	87.2	37.6	107.5
4			√	86.7	87.0	88.1	31.2	93.6
5	√		√	88.7	90.5	88.4	31.4	93.8
6		√	√	88.8	87.7	87.8	31.4	93.7
7	√	√	√	90.8	90.1	88.6	31.4	93.8

B. *Comparative Experimental Analysis of Applicability of YOLO Series*

The fundamental idea of the YOLO series algorithms is to divide the image into fixed-sized grids and make predictions for each grid, thereby achieving object detection. To validate the general applicability of the three proposed modules in the YOLO series algorithms, we conducted comparative experiments on the YOLOv4-YOLOv7 [31-33] algorithms. The experimental results are shown in Table III.

From the data in the table, it can be observed that when detecting on the dataset of eighteen classes used in this paper, the original YOLO series algorithms achieved average precisions of 80.7%, 85.7%, 84.5%, and 86.3% respectively. After integrating the proposed improvement methods into each model, the average precisions were improved by 1.8%, 1.6%, 2.3%, and 4.5% respectively. The experimental results demonstrate that by adding the three proposed modules to the backbone networks of each YOLO algorithm, the average precisions of the models were improved to varying degrees. This is attributed to the fact that the proposed modules refine the feature extraction process, enhance the network's adaptability to multi-scale targets, and bolster the network's robustness in complex backgrounds through structural improvements. Therefore, the proposed improvement methods can be widely applied in various YOLO series algorithms to reduce model complexity, enhance the learning ability and feature extraction capability of the networks, and solve the problems of multi-scale targets, complex backgrounds, and diverse scenes in images captured in power construction sites.

C. *Comparative Experimental Analysis with Other Models*

To validate the advantages of the proposed improved model compared to the current state-of-the-art object detection algorithms, we compared our method with commonly used object detection methods, including Faster-RCNN [34], SSD [35], RetinaNet[36], YOLOv5, TPH-YOLOv5 [37], YOLOv7, and YOLOv8. Using the same dataset and partitioning strategy, we trained each model while keeping the parameters consistent. The experimental results are shown in Table IV.

From Table IV, it can be observed that compared to the current state-of-the-art small object detection algorithm TPH-YOLOv5 and other mainstream algorithms, the proposed algorithm in this paper achieves higher accuracy on most categories. The improved algorithm outperforms Faster R-CNN, SSD, and RetinaNet algorithms, with increases in average precision (mAP) of 11.7%, 28.5%, and 7.1% respectively. Compared to the previous version of YOLOv7, the improved algorithm achieves a 4.5% increase in mAP. The experimental results demonstrate that the improved YOLO model achieves better detection accuracy.

Additionally, to visually demonstrate the superiority of the improved algorithm, this paper provides visual results under different detection algorithms (see Fig. 6). From Fig. 6, it can be observed that when detecting in operation scenes with diverse and complex backgrounds, the Faster R-CNN algorithm shows missing detections and inaccurate bounding box localization. For small-scale object categories, such as the "hook" category in lifting operations and the "aqs_hang" category in elevated work scenarios, RetinaNet, SSD, and the previous version of YOLOv7 algorithms suffer from missing detections. Encouragingly, the model constructed in this paper does not exhibit such issues. This can be attributed to the addition of the GLCI module, which enhances the learning ability of the model, allowing the network to focus on the core information of the features and reduce noise interference from the background. Therefore, this method displays high adaptability in recognizing the clothing and equipment of operators, reducing the occurrence of missed detections, and achieving significant improvements in detection accuracy. For unevenly distributed scale categories, such as the "fence" category in lifting operation scenes and the "protective mask" category in hot work scenarios, other algorithms tend to have missing detections, while the proposed algorithm effectively addresses this issue. This is because the TPD module introduced in this paper enhances the network's learning ability for features at different scales, improving the detection accuracy of multi-scale objects.

TABLE III. APPLICABILITY EXPERIMENTS OF YOLO SERIES ALGORITHMS

Model	AP50/%								
	crane	*person*	*head*	*fence*	*czs*	*hook*	*belt*	*wrong_ belt*	*aqs_hang*
YOLOv4	81.2	84.7	79.8	76.4	80.4	81.5	80.9	81.8	83.3
Improved YOLOv4	81.6	85.2	80.8	75.2	83.3	83.7	81.8	84.9	86.0
YOLOv5	85.8	90.6	83.6	78.5	86.3	91.0	84.6	88.8	89.7
Improved YOLOv5	88.3	92.4	84.7	81.7	88.5	91.9	86.2	89.1	92.0
YOLOv6	83.6	91.2	83.8	75.2	86.2	89.3	81.0	89.3	90.5
Improved YOLOv6	83.7	89.7	88.3	76.1	89.1	90.0	82.3	90.2	92.3
YOLOv7	87.4	95.6	86.3	78.7	85.7	91.1	82.5	83.1	91.4
Improved YOLOv7	91.7	97.8	89.7	88.2	89.8	93.3	87.9	90.7	95.3

Model	AP50/%								mAP50
	aqs_ nohang	*fire operator*	*fire watcher*	*protective face shield*	*fire*	*extinguisher*	*hand_ false*	*phone*	*/%*
YOLOv4	79.9	83.4	83.3	91.8	78.2	81.5	75.9	68.6	80.7

Improved YOLOv4	82.1	83.8	85.1	92.7	79.6	82.3	81.5	73.4	82.5
YOLOv5	88.7	85.3	86.5	91.4	81.3	86.9	80.9	76.4	85.7
Improved YOLOv5	89.5	88.3	89.5	92.1	84.2	87.1	82.6	75.3	87.3
YOLOv6	88.9	85.2	87.2	91.4	81.7	80.2	79.0	72.7	84.5
Improved YOLOv6	88.5	87.5	94.7	93.8	84.5	88.3	81.5	75.9	86.8
YOLOv7	90.1	87.4	84.2	94.7	84.8	89.0	80.1	75.2	86.3
Improved YOLOv7	93.4	90.7	95.2	94.8	90.3	91.8	85.3	77.7	90.8

TABLE IV. COMPARISON OF MODEL PERFORMANCE DIFFERENCES

Model	**AP50/ %**								
	crane	*person*	*head*	*fence*	*czs*	*hook*	*belt*	*wrong_belt*	*aqs_hang*
Faster R-CNN	80.1	84.1	78.3	63.7	79.2	89.4	75.4	81.6	76.9
SSD	63.5	78.7	68.9	55.2	71.8	60.8	63.5	64.2	52.0
RetinaNet	80.4	87.6	86.8	76.5	79.8	89.8	84.2	76.9	83.1
YOLOv5	85.8	90.6	83.6	78.5	86.3	91.0	84.6	88.8	89.7
TPH-YOLOv5	88.4	91.6	89.3	83.3	84.7	91.4	81.3	81.7	90.8
YOLOv7	87.4	95.6	86.3	78.7	85.7	91.1	82.5	83.1	91.4
YOLOv8	91.2	95.5	90.8	86.1	88.6	91.6	92.5	82.1	93.5
Ours	91.7	97.8	89.7	88.2	89.8	93.3	87.9	90.7	95.3
Model	**AP50/ %**								**mAP50**
	aqs_nohang	*fire operator*	*fire watcher*	*protective face shield*	*fire*	*extinguisher*	*hand_false*	*phone*	*/%*
Faster R-CNN	88.3	79.6	65.3	88.5	77.2	84.9	78.6	74.4	79.1
SSD	59.7	71.4	60.3	62.4	61.7	59.1	51.3	53.9	62.3
RetinaNet	89.7	87.3	89.2	84.7	82.3	83.7	79.4	81.2	83.7
YOLOv5	88.7	85.3	86.5	91.4	81.3	86.9	80.9	76.4	85.7
TPH-YOLOv5	90.6	88.4	89.7	91.4	91.3	87.9	81.8	80.3	87.3
YOLOv7	90.1	87.4	84.2	94.7	84.8	89.0	80.1	75.2	86.3
YOLOv8	94.9	87.2	83.1	93.8	87.6	86.3	85.9	84.7	89.1
Ours	93.4	90.7	95.2	94.8	90.3	91.8	85.3	77.7	90.8

D. Dataset Comparison and Model Scalability Experiments

To evaluate the scalability and generalization capabilities of the proposed model, we conducted assessments on two publicly available datasets, SHWD [38] and Pictor-v3 [39], and compared our model with other object detection models. The SHWD dataset focuses on safety helmet detection and contains 7,581 images, including 9,047 instances of workers wearing helmets and 111,514 instances of workers not wearing helmets. The Pictor-v3 dataset primarily focuses on detecting compliance with personal protective equipment (PPE) at construction sites, comprising 1,472 labeled images, which cover various combinations of PPE: 1,209 worker-only instances (W), 2,206 worker instances with helmets (WH), 328 worker instances with vests (WV), and 983 worker instances wearing both helmets and vests (WHV). The evaluation on these two datasets further validates the model's detection capability in different scenarios, as shown in Table V.

As can be seen in Table V, the proposed improvement outperforms other YOLO-based models in both helmet detection and worker PPE compliance detection, with notable improvements in detection accuracy and inference speed. Compared to the TPH-YOLOv5 model, which specializes in small object detection, our improved model demonstrates superior performance in detecting small-scale targets, with significant increases in mAP (0.50) and AP (0.50:0.95) metrics. Additionally, our model also achieves faster inference speeds per image than the original model. Fig. 7 and Fig. 8 show the visual detection results of the improved model and the original model on the SHWD and Pictor-v3 datasets. From the figures, it is evident that the original model exhibits certain false positives and missed detections, particularly when detecting small-scale helmet targets and workers with inconsistent scales. In contrast, the improved YOLOv7 model demonstrates higher precision and robustness in detecting such targets, significantly outperforming the original model, thus further validating the effectiveness of the proposed improvements.

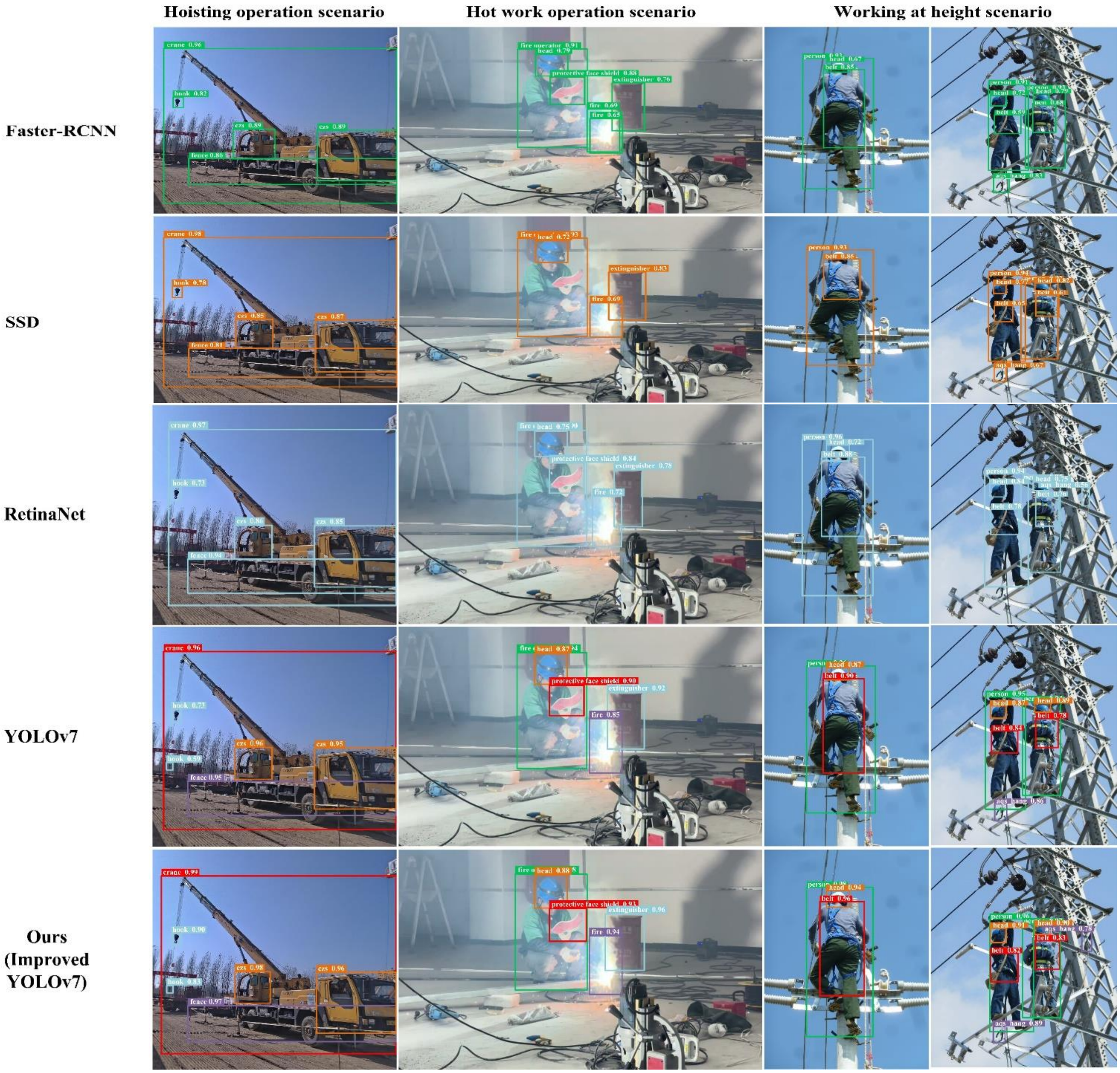
Hoisting operation scenario
Hot work operation scenario
Working at height scenario
Faster-RCNN
SSD
RetinaNet
YOLOv7
Ours
(Improved
YOLOv7)

Fig. 6. Comparison of visual outcomes across diverse models.

TABLE V. PERFORMANCE COMPARISON OF DIFFERENT MODELS ON SHWD AND PICTOR-V3 DATASETS

Model	SHWD			Pictor-v3		
	mAP50 /%	*mAP(0.50:0.95) /%*	*Inference Time /ms*	*mAP50 /%*	*mAP(0.50:0.95) /%*	*Inference Time /ms*
TPH-YOLOv5	86.7	64.2	26.4	88.4	54.8	**19.3**
YOLOv7	87.2	64.6	29.6	89.7	53.9	25.8
YOLOv8	90.8	65.7	41.9	91.3	55.1	33.5
Ours	**91.5**	**66.1**	**26.2**	**92.6**	**56.9**	20.7

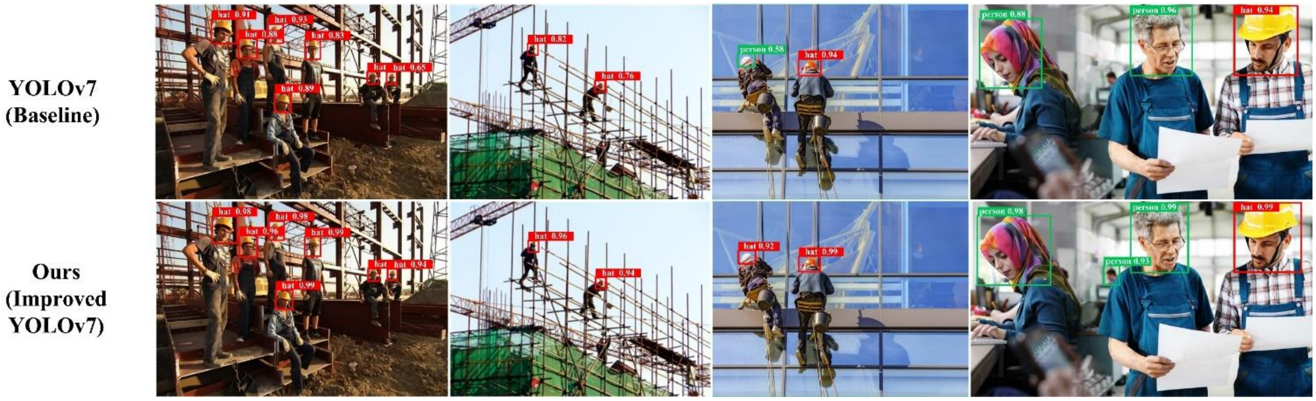

Fig. 7. Comparison of visualization results on the SHWD dataset.

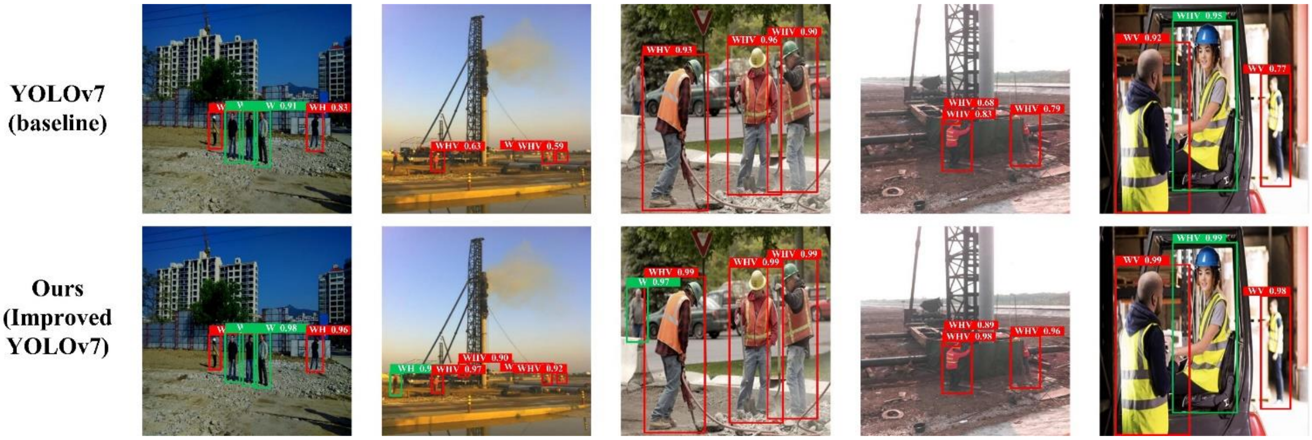

Fig. 8. Comparison of visualization results on the Pictor-v3 dataset.

V. Conclusion

To achieve intelligent compliance recognition in power construction sites, we propose a YOLOv7-based method for detecting personnel behavior that integrates contextual information. This method is designed to address the challenges of high complexity, insufficient scale adaptability, complex image backgrounds, and varying target sizes in existing detection models.

We introduced the GLCI module, which significantly enhances object detection accuracy in complex backgrounds through global and local attention mechanisms. Simultaneously, the TPD module improves the network's ability to learn multi-scale features, leading to better detection performance across varying target scales. Additionally, the C3GhostNetV2 module enhances the model's representational power while reducing computational and parameter complexity. Experimental results demonstrate that the improved YOLOv7 model surpasses the original baseline in detection accuracy, model complexity, and miss rate, showing exceptional adaptability in complex environments. These findings offer effective solutions for object detection tasks in complex scenarios, such as power construction sites, and contribute positively to the advancement of intelligent vision. Furthermore, the dataset comparison and model scalability experiments validate the robustness and generalization capabilities of the proposed model across diverse scenarios. In the future, we aim to enhance the model's robustness across diverse scenarios and lighting conditions, improving adaptability and generalization for deployment on edge devices.

Acknowledgment

Author Contributions: C.B. designed the research and drafted the paper; Z.Y. processed the data; Y.X. debugged the server and configured the experimental environment; Z.L. and D.Y. revised and finalized the paper. All authors have reviewed and approved the final version of the manuscript.

Funding: This research was funded by the Science and Technology Project of State Grid Beijing Electric Power Company (520234230004).

Conflicts of Interest: The authors declare that there are no conflicts of interest.

References

[1] S. Zhang, G. Fu, W. Yin, and P. Gao, “Analysis and prevention of safety helmet accidents based on behavioral safety,” Coal Mine Safety, vol. 45, no. 4, pp. 229–232, 2014.

[2] Y. Wang, Z. Wang, B. Wu, and G. Yang, “Research review of safety helmet wearing detection algorithm in intelligent construction site,” J. Wuhan Univ. Technol., vol. 43, no. 10, pp. 56–62, 2021.

[3] S. Li, H. Ouyang, T. Chen, X. Lu, and Z. Zhao, “Yolo-t: multi-target detection algorithm for transmission lines,” Int. J. Adv. Comput. Sci. Appl., vol. 15, no. 5, 2024..

[4] W. Luo, M. Y. Ahmad Ihsan, M. S. K. Khaizi, and R. Raju, “Hardhat-yolo: a yolov5-based lightweight hardhat-wearing detection algorithm in substation sites,” Int. J. Adv. Comput. Sci. Appl., vol. 15, no. 5, 2024..

[5] H. Nguyen and T. A. Nguyen, "Hybrid vision transformers and cnns for enhanced transmission line segmentation in aerial images," Int. J. Adv. Comput. Sci. Appl., vol. 15, no. 1, 2024..

[6] H. Lei, X. Ge, C. Hao, L. Zhang, and S. Chang, "Identification of dress code of workers in substation based on YOLO v5," Power Syst. Big Data, vol. 24, no. 10, pp. 1–8, 2021.

[7] K. Arai, K. Beppu, Y. Ifuku, and M. Oda, "Method for detecting the appropriateness of wearing a helmet chin strap at construction sites," Int. J. Adv. Comput. Sci. Appl., vol. 15, no. 7, 2024.

[8] B. Ren, Y. Zheng, Y. Wang, S. Sheng, J. Li, and H. Zhang, "Research status and prospect of deep learning in secondary state monitoring of smart substation," in Proc. 2020 Asia Energy and Electrical Engineering Symp. (AEEES), IEEE, 2020, pp. 669–677.

[9] X. Liu, B. Zhang, Y. Fu, and J. Zhu, "Detection on normalization of operating personnel dressing at contaminated sites based on deep learning," J. Safety Sci. Technol., vol. 16, no. 7, pp. 169–175, 2020.

[10] M. Fang, T. Sun, and Z. Shao, "Fast helmet-wearing-condition detection based on improved YOLOv2," Opt. Precision Eng., vol. 27, no. 5, pp. 1196–1205, 2019.

[11] K. Xu and C. Deng, "Research on helmet wear identification based on improved YOLOv3," Laser Optoelectron. Prog., vol. 58, no. 6, pp. 300–307, 2021.

[12] S. Y. Jeon, J. H. Park, S. B. Youn, Y. S. Kim, Y. S. Lee, and J. H. Jeon, "Real-time worker safety management system using deep learning-based video analysis algorithm," Smart Media J., vol. 9, no. 3, pp. 25–30, 2020.

[13] X. Du, Y. Wang, R. Yan, D. Gu, X. Zhang, and T. Lei, "Accurate helmet wearing detection algorithm based on YOLO-ST," J. Shaanxi Univ. Sci. Technol., vol. 40, no. 6, pp. 177–183, 91, 2022.

[14] X. Long, W. Cui, and Z. Zheng, "Safety helmet wearing detection based on deep learning," in Proc. 2019 IEEE 3rd Inf. Technol., Netw., Electron. Autom. Control Conf. (ITNEC), 2019, pp. 2495–2499.

[15] J. Wu, N. Cai, W. Chen, H. Wang, and G. Wang, "Automatic detection of hardhats worn by construction personnel: A deep learning approach and benchmark dataset," Autom. Constr., vol. 106, p. 102894, 2019.

[16] Y. Li, H. Wei, Z. Han, J. Huang, and W. Wang, "Deep learning-based safety helmet detection in engineering management based on convolutional neural networks," Adv. Civ. Eng., vol. 2020, pp. 1–10, 2020.

[17] S. Lang, F. Ventola, and K. Kersting, "DAFNe: a one-stage anchor-free approach for oriented object detection," arXiv preprint arXiv:2109.06148, 2021.

[18] J. Shen, X. Xiong, Y. Li, W. He, P. Li, and X. Zheng, "Detecting safety helmet wearing on construction sites with bounding‐box regression and deep transfer learning," Comput.-Aided Civ. Infrastruct. Eng., vol. 36, no. 2, pp. 180‐196, 2021.

[19] C. Sun, S. Zhang, P. Qu, X. Wu, P. Feng, Z. Tao, J. Zhang, and Y. Wang, "MCA-YOLOV5-Light: A faster, stronger and lighter algorithm for helmet-wearing detection," Appl. Sci., vol. 12, no. 19, p. 9697, 2022.

[20] J. Y. Lee, W. S. Choi, and S. H. Choi, "Verification and performance comparison of CNN-based algorithms for two-step helmet-wearing detection," Expert Syst. Appl., vol. 225, p. 120096, 2023.

[21] S. Yue, Q. Zhang, D. Shao, Y. Fan, and J. Bai, "Safety helmet wearing status detection based on improved boosted random ferns," Multimed. Tools Appl., vol. 81, pp. 16783–16796, 2022.

[22] B. Zhang, C. F. Sun, S. Q. Fang, Y. H. Zhao, and S. Su, "Workshop safety helmet wearing detection model based on SCM-YOLO," Sensors, vol. 22, no. 17, p. 6702, 2022.

[23] S. Bourou, A. Maniatis, D. Kontopoulos, and P. A. Karkazis, "Smart detection system of safety hazards in Industry 5.0," Telecom, vol. 5, no. 1, pp. 1–20, 2023.

[24] A. T. A. Al Daghan, S. Kesh, and A. S. Manek, "A deep learning model for detecting PPE to minimize risk at construction sites," in 2021 IEEE Int. Conf. Electronics, Computing and Communication Technologies (CONECCT), 2021, pp. 1–6.

[25] F. Gong, X. Ji, W. Gong, X. Yuan, and C. Gong, "Deep learning based protective equipment detection on offshore drilling platform," Symmetry, vol. 13, no. 6, p. 954, 2021.

[26] H. Jinqiang, L. Ruihai, L. Hao, L. Yongli, G. Bo, H. Yanpeng, L. Wei, W. Jianrong, and W. Yi, "Visible light image automatic recognition and segmentation method for overhead power line insulators based on YOLO v5 and GrabCut," Southern Power System Technology, vol. 17, no. 06, pp. 128–135, 2023.

[27] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon, "BAM: bottleneck attention module," arXiv preprint arXiv:1807.06514, 2018.

[28] B. Chen and Z. Dang, "Fast PCB defect detection method based on FasterNet backbone network and CBAM attention mechanism integrated with feature fusion module in improved YOLOv7," IEEE Access, 2023.

[29] Y. Tang, K. Han, J. Guo, C. Xu, C. Xu, and Y. Wang, "GhostNetv2: enhance cheap operation with long-range attention," Adv. Neural Inf. Process. Syst., vol. 35, pp. 9969–9982, 2022.

[30] A. H. C. Fong, K. Yoo, M. D. Rosenberg, S. Zhang, C.-S. R. Li, D. Scheinost, R. T. Constable, and M. M. Chun, "Dynamic functional connectivity during task performance and rest predicts individual differences in attention across studies," NeuroImage, vol. 188, pp. 14–25, 2019.

[31] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: optimal speed and accuracy of object detection," arXiv preprint arXiv:2004.10934, 2020.

[32] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, W. Nie, Y. Li, B. Zhang, Y. Liang, L. Zhou, X. Xu, X. Chu, X. Wei, and X. Wei, "YOLOv6: a single-stage object detection framework for industrial applications," arXiv preprint arXiv:2209.02976, 2022.

[33] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 7464-7475.

[34] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," Advances in Neural Information Processing Systems, vol. 28, 2015.

[35] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: single shot multibox detector," in Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I, vol. 14, 2016, pp. 21–37.

[36] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.

[37] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, "TPH-YOLOv5: improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 2778–2788.

[38] M. Gochoo, "Safety helmet wearing dataset," Mendeley Data, 2021, doi: 10.17632/9rcv8mm682.1.

[39] N. D. Nath, A. H. Behzadan, and S. G. Paal, "Deep learning for site safety: Real-time detection of personal protective equipment," Automation in Construction, vol. 112, p. 103085, 2020.